Master 2 Biologie-Informatique/ Bioinformatique 2017-2018

Université Paris Diderot - Paris 7

Etude de la fonction et des mécanismes d'évolution des séquences répétées centromériques chez les Primates

Sarah Kaddah

Tuteur : Loïc Ponger

Structure et Instabilité des Génomes

MNHN - CNRS UMR 7196 / INSERM U1154 - Sorbonne Universités

Remerciements

Merci à Namrod pour toute la partie sur la bibliographie. Retrouvez ses questions FAQ qui ont permis la rédaction de cette partie.
Merci à f-leb, LittleWhite et Metalman pour leurs conseils et la relecture. Merci à ced et jacques_jean pour la correction orthographique et typographique.

Rapport de stage

Sarah Kaddah

1er mars 2018

Résumé

Votre résumé commence ici... ...

Abstract

Abstract begins here... ...

1 Introduction

1.1 Les séquences centromériques

-> info supp sur l'ADN satellite

Le centromère est une structure chromatinienne caractérisé par la présence de CENP-A. Cette protéine, très conservée au cours de l'évolution, est un variant de l'histone H3. Son rôle est de fixer la position du kinétochore par un mécanisme encore peu connu. En effet, le centromère est le site d'assemblage du kinétochore, un ensemble d'ADN et de protéines. Il joue un rôle important durant la division cellulaire chez les eucaryotes en permettant l'attachement du fuseau mitotique pour la ségrégation des chromosomes. Le centromère et les protéines impliquées sont relativement bien conservés. Au contraire, l'ADN sous-jacent est très diversifié et l'organisation varie d'un taxon à l'autre. Cependant, une caractéristique commune est retrouvée chez toute les espèces : de l'ADN centromérique répété en tandem nommé ADN satellite. Ces séquences représentent 5% du génome. Les répétitions s'étendent de 7pb à 3,2kb avec des séquences de 145-180kb le plus souvent.

1.2 L'ADN α-satellites

-> première mise en évidence des AS
->théorie gradient de l'âge
-> travaux sur le gorille à dev

L'ADN satellite chez les Primates est connu sous le nom d'α-satellite. Ces séquences centromériques répétées en tandem, riches en AT, sont issues d'un évènement d'amplification. Un monomère a une longueur de 171pb et il peut être répété des milliers de fois. Les monomères peuvent être répartis en famille selon leur similarité, les séquences ayant un taux d'identité supérieur à 70%. Ces séquences ont soit une organisation monomérique soit une organisation en répétition d'ordre supérieur. Dans le premier cas, les séquences d'une même famille sont répétés en tandem. Dans le deuxème cas, une suite de monomères appartenant à différentes familles forme une unité, qui elle est répétée en tandem. Ces séquences peuvent avoir un site de liaison à la protéine centromérique CENP-B reconnaît la CENP-B box, un motif spécifique de 17pb. Cette protéine, qui reconnaît et se fixe sur l'ADN, serait présente chez de nombreuses familles de primates. La protéine pJα, une protéine peu caractérisée, reconnaît un motif qui remplace la CENP-B box.

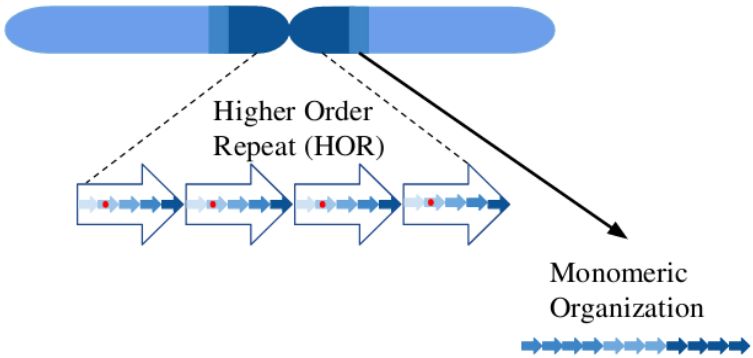

FIGURE 1 – Organisation spatiale des α-satellites. Une couleur correspond à un monomère d'une famille. Les points rouges représentent les sites de fixation à CENP-B.

Les α-satellites ont essentiellement été étudiées chez l'homme. Modèle évolutif avec les centromères en expansion. Une hypothèse concernant l'âge des séquences découle de ces recherches : les séquences les plus récentes apparaissent au coeur du centromères, déplançant les plus anciennes au péricentromère. D'autres études chez le gorille ont été faites. Le rôle des α-satellites est encore mal connu.

1.3 Le sujet de stage

enchaîne sur l'étude chez les cerco, une autre étude de séquençage haut débit

-travaux précédents limités (expliquer pk)
-présentation de mon équipe de leur travail sur les cerco
-la méthode de classification qui a été faite
-Objectif du stage : utilisation de la méthode et comparaison avec des travaux qui ont déjà été faits

Analyser des données de séquençage haut débit en utilisant une méthode de classification identique pour toutes les espèces

2 Matériel et méthode

2.1 Choix des espèces

Les critères de sélections dépendent de la disponibilité des séquences de qualité. Deux espèces du laboratoire, les *Cercopithèques solatus* et *pogonias*, et deux espèces proches, le *Macaca fascicu-*

laris et le *Chlorocebus sabaeus*, sont choisies.

2.2 Méthode de classification

Cette méthode [f.jornod] répartit des séquences α-satellites en familles selon la similarité. La classification est hiérarchique dichotomique. Les séquences sont séparées en fonction de la fréquence des k-mers qui composent les séquences. Cette valeur k est déterminée à 5 à partir des études sur les *Cercopithèques*. Elle décrit la diversité des séquences en optimisant l'usage de la mémoire. Une table de 5-mers est calculée pour toutes les séquences en début de programme.

La classification est suivie d'une double validation des sous-groupes. D'une part la taille du sous-groupe est vérifiée. La taille minimale d'une famille est fixée à 100. Si un groupe atteint 100 séquences, il n'est pas redivisé. D'autre part les deux groupes doivent être distincts. Pour cela le matepair est évalué. C'est la proportion de monomères ayant son plus proche voisin dans le même groupe. Des valeurs matepairs élevées indiquent des sous-groupes bien homogènes et séparés validant la classification tandis qu'un seuil matepair plus faible entraînera plus de classes. Si les matepairs sont au dessus d'un certain seuil, les deux sous-groupes sont ajoutés séparément à la file pour être potentiellement redivisés ultérieurement.En revanche, si au moins une des valeurs de matepair est au dessous de ce seuil, les sous-groupes seront considérés comme formant un seul groupe et le groupe initial est sauvegardé comme une famille unique.

La séparation des séquences se fait de façon itérative en boucle. Chaque tour implique une analyse en composante principale (ACP), une classification hiérarchique et une analyse discriminante linéaire (LDA) si le jeu de données est conséquent. L'ACP est faite sur la table des k-mers pour réduire les dimensions du jeu de données en minimisant la perte d'information et obtenir des variables indépendantes utilisables pour la LDA. Le nombre de composantes est fixé à 1024. Ensuite des distances euclidiennes sont calculées entre toutes les paires de séquences dans l'espace défini par les M premières composantes de l'ACP. Puis la méthode de classification hiérarchique de Ward forme des classes de façon à minimiser l'inertie interclasse. Cependant, si le jeu de données dépasse 110 000 séquences, le calcul des distances devient pesant. La LDA entre alors en jeu. Cette méthode d'apprentissage utilise un sous-jeu de données formé par des séquences tirées aléatoirement. Le modèle construit est appliqué sur toutes les séquences.

2.3 Alignement, consensus et phylogénie

3 Résultat

4 Discussion

5 Conclusion